Theor Appl Genet (1995) 91:534–543

E. Gunel · S. Wearden

Bayesian estimation and testing of gene frequencies

Received: 17 March 1994 / Accepted: 17 January 1995

Abstract This article explains estimation of gene frequencies from a Bayesian viewpoint using prior information. How to obtain Bayes estimators and the highest posterior density credible sets (Bayesian counterpart to classical confidence intervals) for gene frequencies is described. Tests of hypotheses are also discussed. A readily available mathematical application package is used to demonstrate the mathematical computations.

Key words Gene frequencies · Diallel model · Bayesian estimation · Highest posterior credible set · Mathematica®

Introduction

Estimation of gene frequencies has been an important aspect of research not in biology but also in epidemiology and medical genetics. However, in some cases the point estimators which have been in use have been unsatisfactory for a number of reasons. For example, in the multiple alleles situation when there is dominance the classical estimates for the gene frequencies will not necessarily sum to unity and must be adjusted. As for the confidence intervals for gene frequencies, they are only approximate. Furthermore, the standard deviations of the estimates must be approximated, and this can lead to disappointingly wide confidence intervals. On the other hand, Bayesian estimates for the gene frequencies sum to unity, and the credible sets (Bayesian counterpart to classical confidence intervals) are exact.

While the above-mentioned statistical concerns associated with the estimation of gene frequencies can be overcome through the use of Bayesian procedures, even with the Bayesian methods, the computational difficulties still remain. However, recent developments in computing software have made much more tractable many of the mathematical computations that were formerly very tedious. Thus, it is now possible and practical to obtain reliable point and interval estimates of gene frequencies.

The main advantage of the Bayesian analysis is its ability to incorporate non-sample information into the analysis. It is a common-sense approach where we use the observed data and all other non-experimental information to make decisions. In doing so, we can take our previous experience into account explicitly. Most geneticists are well-versed in classical procedures. On the other hand, Bayesian analysis is not commonly used in genetics even though there are circumstances where prior information is available and it would be advantageous to use Bayesian analysis.

The purpose of this paper is to obtain and explain Bayesian estimates of gene frequencies and to demonstrate how the mathematical computations can be performed with Mathematica®, a mathematical application package. Three genetic models will be discussed, and for each will be presented: (1) the traditional method of estimating gene frequency, (2) the Bayesian alternative, and the Mathematica commands and output that provide the solutions.

Bayesian analysis

Suppose we are interested in estimating an unknown numerical quantity θ (possible a vector). We conduct an experiment to obtain information about θ. As a result of the experiment we observe the value x of a random variable X (possible a vector). The probability density (or mass) function $f(x|\theta)$ of X given θ models the uncertainty about X. Classical statistical procedures use $f(x|\theta)$ to make inferences about θ. Bayesian analysis goes one step further and incorporates the non-experimental information about θ into the analysis. The information about θ prior to observing X is called prior information, and it either comes from past experiences involving similar θ or it simply represents our educated

Communicated by D. Van Vleck

E. Gunel (✉) · S. Wearden
Department of Statistics & Computer Science, West Virginia University, Morgantown, W V 26506, USA

opinions. We then specify a prior probability density (or mass) function $\pi(\theta)$, which models our prior knowledge and uncertainties about θ. By using the Bayes paradigm, after observing $X = x$ we obtain the posterior probability density (or mass) function $\pi(\theta|x)$, which models the posterior uncertainty about θ. In Bayesian analysis $\pi(\theta|x)$ plays the central role in making inferences about θ, rather than $f(x|\theta)$ as in classical analysis. For an excellent treatment of the Bayesian theory see Berger (1985).

The diallel model

The measure of many phenotypic variables is controlled by only two genes or alleles, one being received at random from the pair of each parent. Such was the situation first described by Gregor Mendel, and it is this diallel model that will be considered here. Under this model, each member of the population carries only two genes, but in the population itself, there may be many alleles, say alleles P, Q and R, which are carried in combinations of two by members of that population. If these genes have frequencies p, q and r, respectively, then the genotypic frequencies will depend on the relative sizes of p, q and r, and on whether or not there is any survival advantage or sexual selection for any of the alleles.

For procedures described here, it is asumed that in the choice of mates there is no selection due to the genotype, and that after conception all genotypes have an equal chance of survival at least until the time of measurement. Hence, the probability of a genotypic combination depends only on the frequencies of the genes in the population. Such a population is said to be a random mating or panmictic population with the following genotypic frequencies:

Genotypic combination:	PP	PQ	PR	QQ	QR	RR
Genotypic frequency:	p^2	$2pq$	$2pr$	q^2	$2qr$	r^2

where $0 \le p, q, r \le 1$ and $p + q + r = 1$.

Depending on the nature of the trait under such genetic control, it is frequently of interest to research workers to estimate p, q and r but depending on the type of inheritance involved, it may not be possibe to distinguish between two different genotypic combinations. For example, the variable controlled by genotypes PP and PR may receive the same nominal measure. In all cases considered, we assume Hardy-Weinberg equilibrium and no linkage disequilibrium. C.C. Li (1955) discussed the traditional point estimators as well as the problem of setting confidence intervals for the resulting estimates. This paper will show how Bayesian estimation and Mathematica can be used to obtain both point and interval estimates.

Two alleles without dominance

In some cases, each genotype provides a different nominal measure or phenotype for the variable, thereby making it possible to identify the genotypes and obtain estimates of genotypic frequencies from the phenotypic frequencies. For ease of computation, such a situation will be discussed when only two alleles, P and Q, are involved. Hence $r = 0$, and it is of interest to estimate p and $q = 1 - p$. In such a situation, the genotypic combinations and their frequencies are

Genotype and phenotype	PP	PQ	QQ
Phenotypic frequency	p^2	$2pq$	q^2

Let x_1, x_2, x_3 denote the observed number of individuals in groups PP, PQ and QQ, respectively, where $x_1 + x_2 + x_3 = n$. The likelihood function is

$$L(p) = \{n!/x_1!x_2!x_3!\}p^{2x_1}[2p(1-p)]^{x_2}(1-p)^{2x_3}$$

The maximum likelihood estimator of p is $(2x_1 + x_2)/2n$.

The beta prior distribution is a mathematically tractable so-called conjugate prior distribution for p. The uncertainty we have about p can be modeled by using a member of the family of beta prior distributions. Let p have the following beta probability density function,

$$\pi(p) = p^{a-1}(1-p)^{b-1}/B[a,b] \quad \text{if } 0 < p < 1$$

where $a, b > 0$ and $B[a,b] = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the complete beta integral. Symbolically, $p \sim \text{Beta}[a,b]$ and the parameters a and b of the prior distribution are called hyperparameters. We present our prior beliefs about p by using a specific member of the beta family of prior distributions. That is, we use our prior beliefs about p to assess the hyperparameters a and b.

The prior expected value and prior variance of p are given as follows,

$$E(p) = a/(a+b)$$

$$V(p) = ab/[(a+b)^2(a+b+1)].$$

The reciprocal of the prior variance $H(p) = 1/V(p)$ is called the prior precision of p.

Solve $E(p)$ and $V(p)$ for a and b,

$$a = E(p)(a+b)$$

$$b = [1 - E(p)](a+b)$$

where

$$a + b = E(p)[1 - E(p)]H(p) - 1.$$

One way to assess a and b is to specify $E(p)$ and $H(p)$. The posterior distribution of p is again a beta distribution

$$p|(x_1, x_2, x_3) \sim \text{Beta}[2x_1 + x_2 + a,\ x_2 + 2x_3 + b]$$

The posterior distribution models the uncertainty about p after observing x_1, x_2, x_3. Under the quadratic

loss function the Bayes estimator of p is the posterior mean

$$E(p|x_1, x_2, x_3) = (2x_1 + x_2 + a)/(2n + a + b)$$

$$= \frac{2n\dfrac{2x_1 + x_2}{2n} + E(p)U}{2n + U}$$

which is a weighted average of the prior mean and the maximum likelihood estimate of p. A change of two in U can be thought of as being equivalent to a change of one in the sample size n. Consequently, $U = a + b$ can be thought of as being a measure of the amount of prior knowledge about p introduced by the analyst. Small values of $U = a + b$ correspond to little prior knowledge. The posterior mean and the maximum likelihood estimate will be similar when U is close to zero.

The marginal distribution, the Bayesian prior predictive distribution, of (x_1, x_2, x_3) is obtained as the beta mixture of the sampling distribution

$$f(x_1, x_2, x_3|p) = \{n!/x_1!x_2!x_3!\}p^{2x_1 + x_2}(1-p)^{x_2 + 2x_3}2^{x_2}$$

$$\text{if } (x_1, x_2, x_3) \varepsilon A$$

where $A = \{(x_1, x_2, x_3)\colon x_1, x_2, x_3 > 0,\ x_1 + x_2 + x_3 = n\}$. For $(x_1, x_2, x_3) \varepsilon A$ the marginal probability mass function is

$$m(x_1, x_2, x_3) = E^{\pi}[f(x_1, x_2, x_3|p)]$$

$$= \{n!/x_1!x_2!x_3!\}2^{x_2}$$

$$\times B[2x_1 + x_2 + a, x_2 + 2x_3 + b]/B[a, b]$$

where E^{π} denotes the expectation with respect to the prior distribution. Here we have two free variables and the probability mass function expressed in terms of any two of the variables takes the above form. The marginal distribution reflects the plausibility of the prior distribution and can be considered as a likelihood function for the prior distribution.

Other inferences about p are straightforward. To test the null hypothesis $H_0\colon p = p_h$ against $H_1\colon p \neq p_h$ where $0 < p_h < 1$ is a given constant, we would compute the Bayes factor F (H_0) by taking a Beta$[a, b]$ distribution for p under H_1.

$$\mathrm{F}(H_0) = f(x_1, x_2, x_3|p_h)/m(x_1, x_2, x_3)$$

$$= p_h^{2x_1 + x_2}(1 - p_h)^{x_2 + 2x_3}$$

$$\times B[a, b]/B[2x_1 + x_2 + a, x_2 + 2x_3 + b].$$

Bayes factor values of less than one give evidence against the null hypothesis. If we denote the prior probability of H_0 by $\mathrm{P}(H_0)$ then the posterior probability of H_0 is

$$\mathrm{P}(H_0|x_1, x_2, x_3) = \frac{\mathrm{O}(H_0)\,\mathrm{F}(H_0)}{1 + \mathrm{O}(H_0)\,\mathrm{F}(H_0)}$$

where $\mathrm{O}(H_0) = \mathrm{P}(H_0)/(1 - \mathrm{P}(H_0))$ is the prior odds in favor of H_0.

For the mathematics of Bayes factors see Dickey (1971, 1976) and Dickey and Gunel (1978). Gunel and Dickey (1974) discuss Bayes factors for categorical data.

We also can obtain easily a credible set (Bayesian counterpart to classical confidence intervals) for p. A general treatment of credible sets can be found in Berger (1985). If we let $\pi(p|x_1, x_2, x_3)$ denote the posterior probability density function of p, then $100(1-\alpha)\%$ highest posterior density (HPD) credible set for p is the subset C of $\Omega = \{p\colon 0 < p < 1\}$ of the form $\mathrm{C} = \{p \varepsilon \Omega\colon \pi(p|x_1, x_2, x_3) \geq k(\alpha)\}$ where $k(\alpha)$ is the largest constant such that the posterior probability $Pr(p \varepsilon \mathrm{C}|x_1, x_2, x_3) \geq 1 - \alpha$.

Example 1

Let p and q represent the probabilities of having M and N blood-type genes. For most human populations $p > q$, and gene frequency ratio p/q very seldom is greater than two. That is, for most human races $0.5 < p < 0.66$ with high probability. However, for Eskimos and Hindus p is very large, and for Australian aborigines p is very small. Hence, if the interval estimates for p overlap for two populations, genetic similarity for the two populations may be indicated. As a numerical example of Bayesian estimation we consider the blood types of 569 pure Eskimos from East Greenland as reported by Li(1955, page 25).

Blood type

M	MN	N
475	89	5

Suppose before the above data were collected we have had no reason to suspect that p(M gene frequency for Eskimos) was any different from that of most other human races. Hence, prior to observing the data, suppose we felt that $\mathrm{P}(p < 0.5) = \mathrm{P}(p > 0.66) = 0.05$ and that 58% of the Eskimos most probably have the M gene. These prior opinions can be turned into a beta probability density function by setting $E(p) = 0.58 = a/(a + b)$ and by supposing that the upper and lower limits stated for p correspond roughly to the upper and lower 5% points of the beta distribution. Using Mathematica or tables of the percentage points of the beta distribution as in Pearson and Hartley (1958), we find $a = 61$, $b = 44$. For the Beta[61, 44] prior distribution we have $E(p) = 0.58$, $\mathrm{P}(p < 0.5) = 0.0475$, $\mathrm{P}(p > 0.66) = 0.0476$, and these very closely agree with what we want to express as our prior knowledge. The plot of the probability density function of Beta[61, 44] distribution is given in Fig. 1.

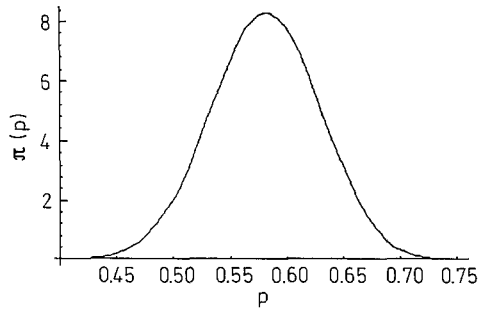


Fig. 1 Graph of the prior p.d.f of p

Bayes procedures are used to obtain point and interval estimates (highest posterior density credible sets) for p, with the computer commands for Mathematica given in bold type. The Mathematica output is given either in the graphs that are presented or as the italic type that follows the bold type for the computer commands.

*We specify the parameters of the beta prior distribution and plot the prior p.d.f

a = 61
b = 44
≪Statistics'ContinuousDistributions'
f[p_] = PDF[BetaDistribution[a, b], p]
plotprior = Plot[f[p], {p, 0.40, 0.75}]

*Even though the prior density function is positive on (0, 1), the function is very close to zero for values of p less than 0.4 and greater than 0.75. Thus, the graph is presented for $0.40 < p < 0.75$

*Enter the observed group frequencies

x_1 = 475
x_2 = 89
x_3 = 5

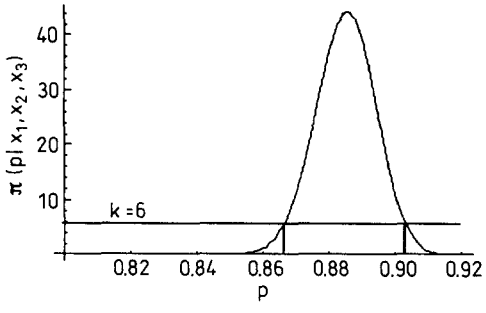
*Compute n and the posterior parameters

n = x_1 + x_2 + x_3
a_2 = 2*x_1 + x_2 + a
b_2 = x_2 + 2*x_3 + b

*Graph the posterior probability density of p

≪Statistics'ContinuousDistributions'
f2[p_] = PDF[BetaDistribution[a_2, b_2], p]
plotposterior = Plot[f2[p], {p, 0.77, 0.86}]

Fig. 2 Graph of the posterior p.d.f of p



*To obtain the estimate of p, compute the posterior mean

m2 = N[(a_2)/(a_2 + b_2)]
0.884956

*Select the value $k = 6$ which will result in small probabilities in the tails and hence a HPD credible set with large coefficient. Now find the roots of $\pi(p|x_1, x_2, x_3) = k$.

FindRoot[f2[p] = = 6, {p, 0.86}]
{p − > 0.866701}

FindRoot[f2[p] = = 6, {p, 0.91}]
{p − > 0.9028}

*Mathematica requires a numerical starting value in search of the root. From the graph we see that the roots of $\pi(p|x_1, x_2, x_3) = 6$ are in the neighborhood of 0.86 and 0.91 and hence we choose these values as starting points in search of the roots. The roots of $\pi(p|x_1, x_2, x_3) = 6$ are 0.866 and 0.902. Now integrate the posterior p.d.f. from 0.866 to 0.902 and obtain the value of $1 - \alpha$. Clearly over (0.866, 0.902) $\pi(p|x_1, x_2, x_3) \geq k = 6$.

F2[p_] = CDF[BetaDistribution[a_2, b_2], p]
coefficient = F2[0.902] − F2[0.866]
0.95296

*Thus (0.866, 0.902) is a 95.29% HPD credible set for p.

Bayes factor computations:

*Suppose we want to test $H_0: p = 0.58$ against $H_1: p \neq 0.58$. Hence set $p_h = 0.58$.

ph = 0.58

*Enter the prior odds $O(H_0)$ in favor of H_0.

oh = 1

*Enter the observed group frequencies.

x_1 = 475
x_2 = 89
x_3 = 5

*Enter the prior distribution parameters.

a = 61
b = 44

*Compute the Bayes Factor.

fh = ph^(2*x_1 + x_2)*(1 − ph)^(x_2 + 2*x_3)*Beta[a, b]/Beta[2*x_1 + x_2 + a, x_2 + 2*x_3 + b]
$8.43876\ 10^{-122}$

*Compute the posterior probability of H_0.

Posterior Probability = oh*fh/(1 + oh*fh)
$8.43876\ 10^{-122}$

*Combine prior and posterior density plots into one.

Show[plotprior, plotposterior, PlotRange- > {{0, 1}, {0, 50}}]

By using Bayesian analysis, in the light of the data we revise our opinions about p. Our point estimate of p is now the posterior mean 0.88 and the posterior distribution of p is Beta$[2x_1 + x_2 + a,\ x_2 + 2x_3 + b]$ = Beta

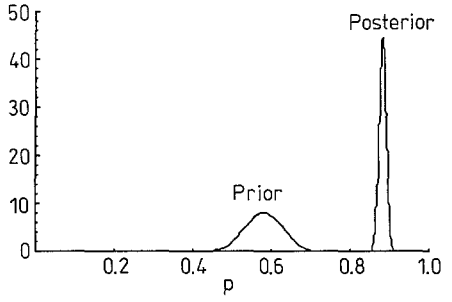


Fig. 3 Combined graphs of the the prior and posterior p.d.f's

[1100, 143]. The posterior probability of the null hypothesis $H_0: p = 0.58$ is 8.43×10^{-122}, which is extremely small. From the HPD credible set (0.866, 0.902) we deduce that with probability 0.95, 86.6% to 90.2% of the Eskimos have the M gene, which indicates that the frequency of the M gene in Eskimos is considerably larger than that of most other races. The Beta[1100, 143] distribution can be used as prior distribution for p prior to observing future data concerning the blood types of Eskimos from East Greenland.

Two alleles with dominance

This situation is similar to that just discussed except both the PP and PQ genotype show the same dominant character, meaning the nominal measure for the variable is the same for both. Hence the situation is

Phenotype	Dominant (PP or PQ)	Recessive(QQ)
Phenotypic frequency	$p^2 + 2pq$	q^2

Let x_1, x_2, where $x_1 + x_2 = n$, denote the number of individuals in groups A and B, respectively. The likelihood function is

$$L(p) = C^n_{x_1}[p^2 + 2p(1-p)]^{x_1}[(1-p)^2]^{x_2}$$

where $C^n_{x_1} = n!/x_1!(n-x_1)!$. The maximum likelihood estimator of p is $1-(x_2/n)^{1/2}$.

Again we model the prior uncertainty about p by a beta distribution with parameters a and b. Using the expansion

$$[p^2 + 2p(1-p)]^{x_1} = \sum_{i=0}^{x_1} C^{x_1}_i p^{2i}[2p(1-p)]^{x_1-i}$$

the posterior probability density function of p can be shown to be a weighted average of beta densities.

$$\pi(p|x_1, x_2) = \sum_{i=0}^{x_1} w(i) f_{\text{beta}}(p|x_1 + a + i, x_1 + 2x_2 + b - i)$$

where $f_{\text{beta}}(p|x_1 + a + i, x_1 + 2x_2 + b - i)$ denotes a beta density with parameters $x_1 + a + i$ and $x_1 + 2x_2 + b - i$ and

$$w(i) = \frac{C^{x_1}_i 2^{x_1 - i} B[x_1 + a + i, x_1 + 2x_2 + b - i]}{\sum_{i=0}^{x_1} C^{x_1}_i 2^{x_1 - i} B[x_1 + a + i, x_1 + 2x_2 + b - i]}$$

The posterior mean of p is then equal to the weighted average of $(x_1 + 1)$ means

$$E[p|x_1, x_2] = \sum_{i=0}^{x_1} w(i)\{(x_1 + a + i)/(2n + a + b)\}.$$

The Bayes factor for testing $H_0: p = p_h$ against $H_1: p \neq p_h$ where $0 < p_h < 1$ is a given constant is given by

$$\sum_{i=0}^{x_1} w(i) p_h^{x_1 + i}(1 - p_h)^{x_1 + 2x_2 - i}$$

$$\times B[a, b]/B[x_1 + a + i, x_1 + 2x_2 + b - i]$$

We now give an example and illustrate how point estimates and a HPD credible set for p can be obtained using Mathematica.

Example 2

R.A. Fisher (1954) gave an example of dominance/recessive inheritance in pea plants. In a random sample of the offspring from a mating population, 60 plants were identified by physical appearance as showing the dominant characteristic (having genotype PP or PQ) or showing the recessive characteristic (having genotype QQ)

Dominant characteristic	Recessive characteristic
36	24

Suppose prior to observing the data, we felt that p is most likely to be in the neighborhood of 0.5 and that $P(p < 0.3) = P(p > 0.7) = 0.01$. By means of the same procedure as in Example 1, these prior opinions can be represented by a beta prior probability density function with parameters $a = b = 15$. For the Beta[15, 15] prior distribution we have $E(p) = 0.5$, $P(p < 0.3) = P(p > 0.66) = 0.011$, and these agree with what we want to express as our prior knowledge.

*We specify the parameters of the beta prior distribution and plot the prior p.d.f

```
a = 15
b = 15
<< Statistics'ContinuousDistributions'
f[p_] = PDF[BetaDistribution[a, b], p]
plotprior = Plot[f[p], {p, 0.20, 0.80}]
```

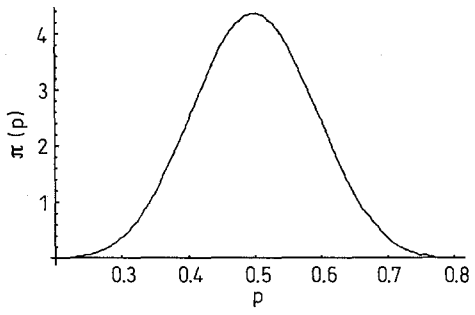


Fig. 4 Graph of the prior p.d.f of p

*Enter the observed group frequencies
$x_1 = 36$
$x_2 = 24$

*Compute $w(i)$
$n = x_1 + x_2$
$w_1[i\_] = (2\hat{}(x_1 - i))*$ Binomial$[x_1, i]$
***Beta$[x_1 + a + i, x_1 + 2*x_2 + b - i]$**
wdot = Sum$[w_1[i], \{i, 0, x_1\}]$
$w[i\_] = w_1[i]$/wdot

*Compute and graph the posterior p.d.f of p
≪Statistics'ContinuousDistributions'
f[p_, i_] = PDF[BetaDistribution
$[x_1 + a + i, x_1 + 2*x_2 + b - i], p]$
$f_2[p\_] =$ Sum$[w[i]*f[p, i], \{i, 0, x_1\}]$
plotposterior = Plot$[f_2[p], \{p, 0.2, 0.6\}]$

*Examine the graph of the posterior p.d.f and realize that a $k = 0.5$ value will result in a HPD credible set with $1 - \alpha \geq 0.90$. Now find the roots of $\pi(p|x_1, x_2) = k$.

FindRoot[f2[p] = = 0.5, {p, 0.3}]
{p− >0.295843}

FindRoot[f2[p] = 0.5, {p, 0.5}]
{p −>0.508346}

*We see that the roots $\pi(p|x_1, x_2) = k$ are 0.295 and 0.508. Now integrate the posterior p.d.f from 0.295 to 0.508 and obtain the value of $1 - \alpha$. Clearly over (0.295, 0.508), $\pi(p|x_1, x_2) \geq k$

G[p_, i_] = CDF[BetaDistribution
$[x_1 + a + i, x_1 + 2*x_2 + b - i], p]$
G2[p_] = Sum[w[i]*G[p, i], {i, 0, x_1}]
Coefficient = G2[0.508] − G2[0.295]
0.984586

*Thus (0.295, 0.508) is a 98.45% HPD credible set for p.

*Compute the Bayes estimate of p
m2 = N[Sum[w[i]*(x_1 + a + i)/(2*n + a + b), {i, 0,x_1}]
0.400284

*For comparison compute the maximum likelihood estimate of p
mle = N[1 − Sqrt[x_2/n]]
0.367544

*Combine prior and posterior density plots into one.
Show[plotprior, plotposterior, PlotRange-> {{0, 1}, {0, 10}}]

After observing the data, our point estimate of p now is the posterior mean 0.40. The maximum likelihood estimate of p is 0.367. From the HPD credible set we have $P(0.295 < p < 0.508|x_1, x_2) = 0.98$, that is in the light of the data and our prior beliefs, p is between 0.295 and 0.508 with probability 0.98.

Multiple alleles

The inheritance associated with the blood types has become the classic example of a genetic variable under the control of multiple alleles. It is such a familiar example that multiple alleles can best be explained with regard to it. For the blood types, the phenotypes and their frequencies with random mating and Hardy-Weinberg conditions are

Phenotype	Type A (PP or PR)	Type B (QQ or QR)
Frequency	$p^2 + 2pr$	$q^2 + 2qr$

Type AB (PQ)	Type O(RR)
$2pq$	r^2

Fig. 5 Graph of the posterior p.d.f of p

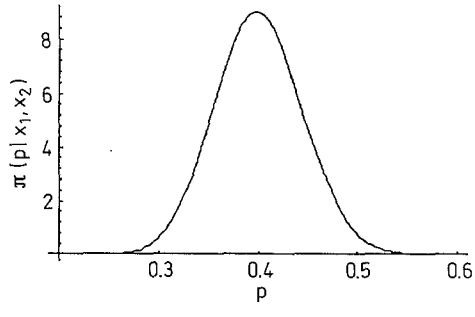
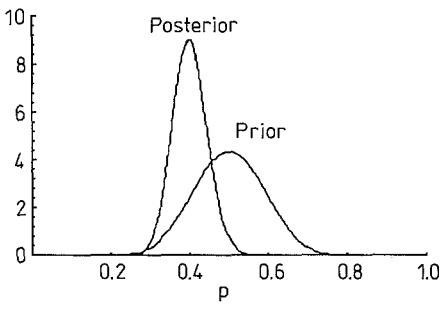


Fig. 6 Combined graphs of the prior and posterior p.d.f's



For a random sample of n unrelated individuals the likelihood function is given by the multinomial law.

$$L(p,q,r) = \{n!/x_1!x_2!x_3!x_4!)\}\,[p^2+2pr]^{x_1}[q^2+2qr]^{x_2}$$

$$[2pq]^{x_3}r^{2x_4}$$

where x_1, x_2, x_3, x_4 $(x_1+x_2+x_3+x_4=n)$ denote the observed number of individuals in groups A, B, AB and O, respectively.

The maximum likelihood estimators of p and q are obtained by solving the following two equations simultaneously for p and q (Stevens 1938);

$$[x_3/p] + [2x_1(1-p-q)/(p^2+2p(1-p-q)]$$

$$-[2x_2q/(q^2+2q(1-p-q)] - [2x_4/(1-p-q)] = 0$$

$$[x_3/q] + [2x_2(1-p-q)/(q^2+2q(1-p-q)]$$

$$-[2x_1p/(p^2+2p(1-p-q)] - [2x_4/(1-p-q)] = 0$$

The equations are nonlinear in p and q and can be solved iteratively via Newton's method by taking

$$p_0 = 1-[(x_2+x_4)/n]^{1/2} \quad q_0 = 1-[(x_1+x_4)/n]^{1/2}$$

as the initial values of p and q.

The estimators p_0, q_0 and $r_0 = [x_4/n]^{1/2}$ have long been used as the estimators of the gene probabilities because they are easy to compute. We will call them classical estimators. The sum $p_0+q_0+r_0$ does not equal one as it should. Thus, these estimates are usually adjusted so that they sum to unity. Bernstein's (1938) method (see Li 1955) of adjustment is to compute

$$p_0' = p_0(1+0.5d),\ q_0' = q_0(1+0.5d),$$

$$r_0' = (r_0+0.5d)(1+0.5d)$$

where $d = 1-(p_0+q_0+r_0)$. If the sum $p_0'+q_0'+r_0' = 1-0.5d^2$ is then appreciably different from one, repeat the process using the new deviation $d' = 0.25d^2$. Usually one adjustment is sufficient to obtain estimates that sum close to unity.

The Dirichlet prior distribution is a mathematically tractable so-called conjugate prior distribution for (p,q,r). The uncertainty about (p,q,r) can be modeled using a member of the family of Dirichlet prior distributions. Let (p,q,r) have a Dirichlet prior distribution with parameters (a,b,c) (Wilks 1962). Since $p+q+r=1$, there are two free variables, and the probability density function of any two takes the identical form

$$\pi(p,q,r) = \{1/B[a,b,c]\}\,p^{a-1}q^{b-1}r^{c-1} \text{ if } (p,q,r)\varepsilon S$$

where $a,b,c>0$, S is the probability simplex $\{p,q,r): p,q,r>0,\ p+q+r=1\}$ and $B[a,b,c] = \Gamma(a)\,\Gamma(b)\,\Gamma(c)/\Gamma(a+b+c)$ is the complete Dirchlet integral. The marginal prior probability distribution of p and q are Beta $[a, b+c]$ and Beta$[b, a+c]$, respectively, and covariance between p and q is $\mathrm{Cov}(p,q) = -ab/[(a+b+c)^2(a+b+c+1)]$. Small values of the sum $a+b+c$ correspond to little prior knowledge about (p,q,r). If we take $a=4, b=6, c=10$, this can be thought of as stating that we would have 4, 6, 10 alleles of types P, Q, R, respectively, in a random sample of 20 observations. On the other hand, if we take $a=16$, $b=24$, $c=40$, we would have the same expected values for p,q,r as in the previous case but now the prior variances of p,q,r would be smaller. In the latter case, we are expressing opinions about p,q,r with more precision. If we have no information about (p,q,r), then this can be expressed by taking a so-called non-informative prior with $a=b=c=0$.

Using the likelihood function, the Dirichlet prior and binomial expansions for $[p^2+2pr]^{x_1}$ and $[q^2+2qr]^{x_2}$, we obtain the following joint posterior probability density for (p,q,r)

$$\pi(p,q,r|x_1,x_2,x_3,x_4)$$

$$= \sum_{i=0}^{x_1}\sum_{j=0}^{x_2} w(i,j)\,f_{\text{Dirichlet}}(p,q,r|a_i', b_j', c_{ij}')$$

where $a_i' = x_1+x_3+i+a$, $b_j' = x_2+x_3+j+b$, $c_{ij}' = x_1+x_2+2x_4-i-j+c$,

$$w(i,j) = \frac{C_i^{x_1}C_j^{x_2}2^{x_1+x_2-i-j}B[a_i',b_j',c_{ij}']}{\sum_{i=0}^{x_1}\sum_{j=0}^{x_2} C_i^{x_1}C_j^{x_2}2^{x_1+x_2-i-j}B[a_i',b_j',c_{ij}']}$$

and $f_{\text{Dirichlet}}(p,q,r|a_i',b_j',c_{ij}')$ denotes the Dirichlet density with parameters a_i', b_j', c_{ij}'. The joint posterior probability density function of (p,q,r) is a weighted average of Dirichlet probability density functions. Moreover, the marginal posterior density functions of p,q and r are weighted averages of beta densities.

$$\pi(\theta|x_1,x_2,x_3,x_4) = \sum_{i=0}^{x_1}\sum_{j=0}^{x_2} w(i,j)\,f_{\text{beta}}(\theta|\alpha_{ij},\beta_{ij})$$

where for $\theta=p$ we have $\alpha_{ij}=a_i'$, $\beta_{ij}=b_j'+c_{ij}'$; for $\theta=q$ we have $\alpha_{ij}=b_j'$, $\beta_{ij}=a_i'+c_{ij}'$ and for $\theta=r$ we have $\alpha_{ij}=c_{ij}'$, $\beta_{ij}=a_i'+b_j'$.

The posterior means of p, q and r are

$$E[\theta|x_1,x_2,x_3,x_4] = \sum_{i=0}^{x_1}\sum_{j=0}^{x_2} w(i,j)\,[h_{ij}/(2n+a+b+c)]$$

where for $\theta=p,q,r$ we have $h_{ij}=a_i', b_j', c_{ij}'$ respectively.

The Bayes factor for testing $H_0: (p,q,r) = (p_h,q_h,r_h)$ against $H_1: (p,q,r) \neq (p_h,q_h,r_h)$ where $0<p_h,q_h,r_h$

$< 1, p_h + q_h + r_h = 1$ are known constants is given by

$$\sum_{i=0}^{x_1}\sum_{j=0}^{x_2} w(i,j)p_h^{a'_i-a}q_h^{b'_j-b}r_h^{c'_{ij}-c}$$

$$\times \mathrm{B}[a,b,c]/\mathrm{B}[a'_i,b'_j,c'_{ij}]$$

The Bayes factors for tests involving individual parameters separately, $H_0: p = p_h$ against $H_1: p \neq p_h$ $H_0: q = q_h$ against $H_1: q \neq q_h$ and $H_0: r = r_h$ against $H_1: r \neq r_h$, are given respectively as follows.

$$\sum_{i=0}^{x_1}\sum_{j=0}^{x_2} w(i,j)p_h^{a'_i-a}(1-p_h)^{b'_j+c'_{ij}-b-c}$$

$$\times \mathrm{B}[a,b+c]/\mathrm{B}[a'_i,b'_j+c'_{ij}]$$

$$\sum_{i=0}^{x_1}\sum_{j=0}^{x_2} w(i,j)q_h^{b'_j-b}(1-q_h)^{a'_i+c'_{ij}-a-c}$$

$$\times \mathrm{B}[b,a+c]/\mathrm{B}[b'_j,a'_i+c'_{ij}]$$

$$\sum_{i=0}^{x_1}\sum_{j=0}^{x_2} w(i,j)r_h^{c'_{ij}-c}(1-r_h)^{a'_i+b'_j-a-b}$$

$$\times \mathrm{B}[c,a+b]/\mathrm{B}[c'_{ij},a'_i+b'_j].$$

Since B[0, 0, 0] or B[0, 0] are undefined, when we are testing the hypothesis, the non-informative prior is expressed by taking $a = b = c = 0.5$.

Example 3

Consider the following data on the distribution of blood groups

Blood group			
A	B	AB	O
30	6	1	34

*We set $a = b = c = 0$ and take the non-informative prior.

```
a = 0
b = 0
c = 0
```

*Enter the observed frequencies

```
x1 = 30
x2 = 6
x3 = 1
x4 = 34
```

*Compute $w(i, j)$

```
n = x1 + x2 + x3 + x4
w1[i_, j_] = (2^(x1 + x2 - i - j))*Binomial[x1, i]*
Binomial[x2, j]*
Gamma[x1 + x3 + a + i]*Gamma[x2 + x3 + b + j]*
Gamma[x1 + x2 + 2*x4 - i - j + c]/
Gamma[2*n + a + b + c]
wdot = Sum[w1[i, j], {i, 0, x1}, {j, 0, x2}]
w[i_, j_] = w1[i, j]/wdot
```

*Compute the marginal posterior p.d.f of p and graph it

```
f[i_, j_] = (p^(x1 + x3 + i + a - 1))*
  (q^(x2 + x3 + j + b - 1))*
((1 - p - q)^(x1 + x2 + 2*x4 - i - j + c - 1))*
Gamma[x1 + x3 + a + i]*
Gamma[x2 + x3 + b + j]*
Gamma[x1 + x2 + 2*x4 - i - j + c]/
Gamma[2*n + a + b + c]
f2[p_, q_] = Sum[w1[i, j]*f[i, j], {i, 0, x1}, {j, 0, x2}]
<<Statistics'ContinuousDistributions'
g21[p_, i_] = PDF[BetaDistribution[x1 + x3 + i + a,
2*n + a + b + c - x1 - x3 - i - a], p]
h21[p_, j_] = Sum[w[i, j]*g21[p, i], {i, 0, x1}]
f21[p_] = Sum[h21[p, j], {j, 0, x2}]
Plot[f21[p], {p, 0.05, 0.5}]
```

*Obtain a HPD credible set for p

```
FindRoot[f21[p] == 1, {p, 0.17}]
{p -> 0.168691}
FindRoot[f21[p] == 1, {p, 0.35}]
{p -> 0.337229}
G21[p_, i_] = CDF[BetaDistribution[x1 + x3 + i + a,
2*n + a + b + c - x1 - x3 - i - a], p]
H21[p_, j_] = Sum[w[i, j]*G21[p, i], {i, 0, x1}]
F21[p_] = Sum[H21[p, j], {j, 0, x2}]
Coefficient1 = F21[0.3372] - F21[0.1686]
0.969616
```

***Thus (0.168. 0.337) is a 96.96% HPD credible set for p

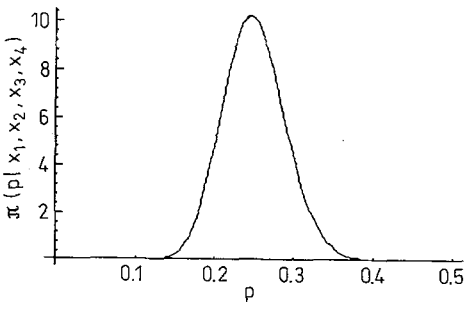
*Compute and graph the posterior p.d.f of q

```
g22[q_, j_] = PDF[BetaDistribution[x2 + x3 + j + b,
2*n + a + b + c - x2 - x3 - j - b], q]
h22[q_, i_] = Sum[w[i, j]*g22[q, j], {j, 0, x2}]
f22[q_] = Sum[h22[q, i], {i, 0, x1}]
Plot[f22[q], {q, 0, 0.14}]
```

*Obtain a HPD credible set for q

```
FindRoot[f22[q] == 2, {q, 0.02}]
{q -> 0.015601}
FindRoot[f22[q] == 2, {q, 0.1}]
{q -> 0.0940032}
```

Fig. 7 Graph of the posterior p.d.f of p



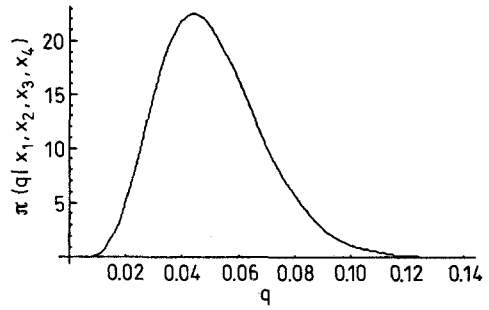


Fig. 8 Graph of the posterior p.d.f of q

G22[q_, j_] = CDF[BetaDistribution[$x_2 + x_3$ + j + b, 2*n + a + b + c − x_2 − x_3 − j − b], q]
H22[q_, i_] = Sum[w[i, j]*G22[q, j], {j, 0, x_2}]
F22[q_] = Sum[H22[q, i], {i, 0, x_1}]
Coefficent2 = F22[0.094] − F22[0.015]
0.97224
***Thus (0.015, 0.094) is a 97.22% HPD credible set for q

*Compute and graph the posterior p.d.f of r
g23[r_, i_, j_] = PDF[BetaDistribution [$x_1 + x_2$ + 2*x_4 − i − j + c, 2*n + a + b + c − x_1 − x_2 − 2*x_4 + i + j − c], r]
f23[r_] = Sum[w[i, j]* g23[r, i, j], {i, 0, x_1}, {j, 0, x_2}]
Plot[f23[r], {r, 0.5, 0.85}]

*Obtain a HPD credible set for r
FindRoot[f23[r] = = 1, {r, 0.6}]
{r −> 0.608891}
FindRoot[f23[r] = = 0.5, {r, 0.77}]
{r −> 0.795488}
G23[r_, i_, j_] = CDF[BetaDistribution [$x_1 + x_2$ + 2*x_4 − i − j + c, 2*n + a + b + c − x_1 − x_2 − 2*x_4 + i + j − c], r]
F23[r_] = Sum[w[i, j]*G23[r, i, j], {i, 0, x_1}, {j, 0, x_2}]
Coefficent3 = F23[0.795] − F23[0.608]
0.976016
***Thus (0.608, 0.795) is a 97.60% HPD credible set for r

*Compute the Bayes estimate of p, q and r
m2p = N[Sum[w[i, j]* ($x_1 + x_3$ + i + a)/ (2*n + a + b + c), {i, 0, x_1}, {j, 0, x_2}]]
0.250572
m2q = N[Sum[w[i, j]* ($x_2 + x_3$ + j + b)/ (2*n + a + b + c), {i, 0, x_1}, {j, 0, x_2}]]
0.0507855
m2r = N[Sum[w[i, j]* ($x_1 + x_2$ + 2*x_4 − i − j + c)/ (2*n + a + b + c), {i, 0, x_1}, {j, 0, x_2}]]
0.69864
***Thus the Bayes estimates of p, q and r are 0.2505, 0.0507, 0.6986, respectively.

*Compute the estimates p_0, q_0 and r_0
p_0 = N[1 − Sqrt[($x_2 + x_4$)/n]]
0.249413
q_0 = N[1 − Sqrt[$x_1 + x_4$)/n]]
0.0505747
r_0 = N[Sqrt[x_4/n]]
0.692007
***The estimates p_0, q_0 and r_0 are 0.2494, 0.0505, 0.6920, respectively.

*Compute the adjusted estimates of p, q and r
sumcl = $p_0 + q_0 + r_0$
0.991995
d = 1.0 − sumcl
0.00800528
adjpo = p_0*(1 + 0.5*d)
0.250412
adjqo = q_0*(1 + 0.5*d)
0.0507771
adjro = (r_0 + 0.5*d)*(1.0 + 0.5*d)
0.698795
sumcl2 = adjpo + adjqo + adjro
0.999984
***Adjusted estimates of p, q and r are 0.2504, 0.0507 and 0.6987, respectively.

*Find the MLE of p and q
g1[p_, q_] = x_3/p + x_1*2*(1 − p − q)/(p^2 + 2*p* (1 − p − q)) − (2*x_2*q)/(q^2 + 2*q* (1 − p − q)) − (2*x_4)/(1 − p − q)

g2[p_, q_] = x_3/q − (2*x_1*p)/(p^2 + 2*p* (1 − p − q)) + x_2*2*(1 − p − q)/(q^2 + 2*q* (1 − p − q)) − (2*x_4)/(1 − p − q)

FindRoot[{g1[p, q] = = 0, g2[p, q] = = 0}, {p, p_0, 0, 1}, {q, q_0, 0, 1}]
{p −> 0.250411, q −> 0.0507771}
***Thus the mle of p, q and r are 0.2504, 0.0507 and 0.6989.

We now summarize the numerical results in Table 1.

In this example we have taken the prior distribution to be non-informative with $a = b = c = 0$. The HPD credible sets and Bayes estimates can be obtained for any other values of a, b and c that reflect the prior information about (p, q, r). We have intentionally chosen $a = b = c = 0$ to show that in the above example, the Bayes estimate posterior mean, and the maximum likelihood estimate are in close agreement. This also

Fig. 9 Graph of the posterior p.d.f of r

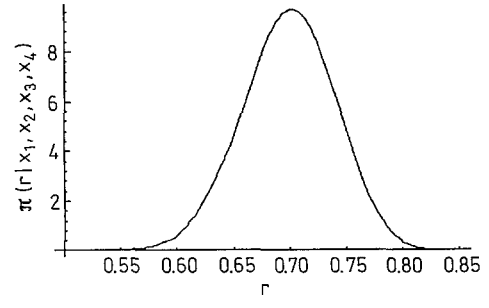


Table 1 Estimates of p, q and r

Parameter	Bayes	97% HPD Credible set	Classical	Classical adjusted	MLE
p	0.2505	(0.168, 0.337)	0.2494	0.2504	0.2504
q	0.0507	(0.015, 0.094)	0.0505	0.0507	0.0507
r	0.6986	(0.608, 0.795)	0.6920	0.6987	0.6989

would have been the case if we had taken a non-informative prior ($a = b = 0$) in Examples 1 and 2.

In the multiple alleles situation when there is dominance the classical estimates for the gene frequencies do not necessarily sum to unity and must be adjusted. The confidence intervals for gene frequencies are approximate. On the other hand, Bayesian estimates for the gene frequencies sum to one and the credible sets are exact.

Acknowledgements We would like to thank the editor and the referees for helpful comments.

References

Berger JO (1985) Statistical decision theory and bayesian analysis, 2nd edn. Springer, Berlin Heidelberg New York

Bernstein F (1938) Uber die Erblichkeit der Blutgrupen. Ibid 54:400–426

Dickey JM (1971) The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann Math Stat 42:203–223

Dickey JM (1976) Approximate posterior distributions. J Am Stat Assoc 71:680–689

Dickey JM, Gunel E (1978) Bayes factors from mixed probabilities. J R Stat Soc Ser, B 40:43–46.

Fisher RA (1954) Statistical methods and scientific inference, 12th edn. Oliver and Boyd, Edinburgh

Gunel E, Dickey JM (1974) Bayes factors for independence in contingency tables. Biometrika 61:545–557

Li CC (1955) Population genetics. The University of Chicago Press, Chicago

Pearson ES, Hartley HO (1958) Biometrika tables for statisticians, vol 1. Cambridge University Press, Cambridge

Stevens WL (1938) Estimation of blood group gene frequencies. Ann Eugen 8:362:375

Wilks SS (1962) Mathematical statistics. John Wiley, New York

Wolfram S (1991) Mathematica, a system for doing mathematics by computer, 2nd edn. Addison-Wesley, Redwood City